Development of Multi-parameter Marker Assays

Lisa M. McShane, Ph.D.

Biometric Research Branch
Division of Cancer Treatment and Diagnosis
National Cancer Institute

February 23, 2008

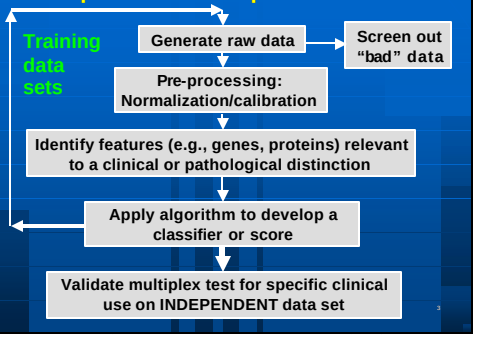
Multi-parameter (multiplex) Assay

Simultaneous measurement of many analytes or characteristics

- Gene expression microarrays
- Multiplex RT-PCR
- SNP chips
- Micro-bead assays
- Multiplex ELISA
- Multiplex proteomics . . .

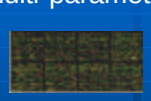
(Much of this talk would apply to situation of many single-analyte tests as well.)

Development of Multiplex Marker Test

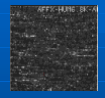


Generate Raw Data

Quantification of pattern output by multi-parameter assay



2-color cDNA array



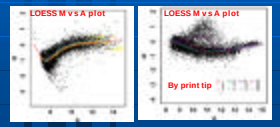
Affymetrix array

Serum proteomics spectral plot

Gene Expression cDNA Microarrays: Pre-process image data ⇒ gene-level data

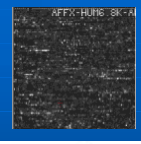
Ratios or 2-channel signals models

Normalization:
- Simple (e.g., median-based)
- Complex (e.g. intensitybased



Intensity-dependent normalization: Yang et al, *Nucl Acids Res 2002*)

Gene Expression Affymetrix Microarrays: Pre-process image data ⇒ gene-level data



MAS-GCOS-AGCC (Affymetrix)
anti-log of
Tukey biweight average of
adjusted log(PMij -IMij)

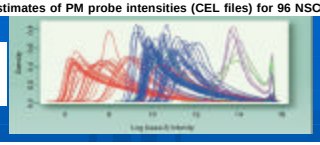
Li & Wong dChip – *PNAS* 2000, *Genome Biology* 2001
MBEIi = ?i or ?i* estimated from
PMij -MMij = ?i f j +ej or
PMij = ?i + ?i* f j´

Irizarry et al – *Biostat & NAR* 2003
RMAi = ei estimated from
T(PMij) = ei + aj +eij
Also GC-RMA – Wu et al – *JASA* 2004

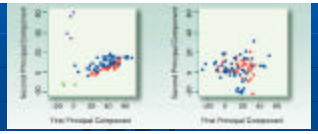
Batch Effects in Gene Expression Data

Density estimates of PM probe intensities (CEL files) for 96 NSCLC samples

Red = batch 1
Blue = batch 2
Purple & Green = outliers?



PCA plots after RMA pre-processing with and without outlier CEL files

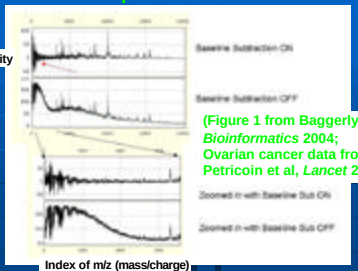


(Figure 1 from Owzar et al, *Clinical Cancer Research* 2008 using data from Beer et al., Nature Medicine 2002)

SELDI-TOF Mass Spectrometry

Process spectra Þ baseline subtraction and peak identification

Intensity



(Figure 1 from Baggerly et al, *Bioinformatics* 2004; Ovarian cancer data from Petricoin et al, *Lancet* 2002)

Index of m/z (mass/charge)

Identification of Features "Informative" for Clinical Outcome or Characteristic

- Gene(s) whose expression correlates with survival
- Protein(s) whose presence is associated with cancer
- SNP(s) whose presence is associated with favorable or toxic response to drug . . .

Informative Feature List Instability

- Multiple testing issues
  - 10,000 non-informative features each tested at 0.05 level of significance will produce 500 false positives
  - Typically use smaller testing level (e.g., 0.001) or more sophisticated procedures
- Size of list dependent on stringency of multiple testing corrections
- Low power under stringent multiple testing corrections
- Co-regulation of genes

Classifier or Multivariate Score

- Link multiplex marker measurements to clinical outcome or characteristic
- Function that associates a specimen with a class or assigns a continuous score based on inputted feature measurements
- Most scores eventually subject to cutpoints for clinical decision-making

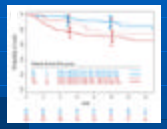
(Focus here on classifier building.)

Multiplex Marker Output

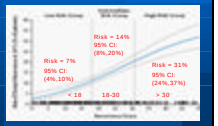
Link to clinical outcome

RISK SCORE or CLASSIFIER

Clinical Test



Inform clinical decision



Buyse et al., *JNCI*, 2006

Mammaprint
Prognostic/predictive?

(Figure 4 from Paik et al., *N Engl J Med*, Dec. 2004)

Oncotype DX
Prognostic/predictive?

Feature List ? Classifier

- Clustering method applied to feature set does not rigorously define a classifier (e.g., see Lusa et al, *JNCI* 2007 discussion of breast cancer subtypes)
  - Results differ by clustering technique
  - Results sensitive to data normalization & centering
  - Results dependent on set of samples to which clustering methods are applied
  - Assignment of clusters to outcome class?
- Classifiers with similar performance may be developed from substantially different feature lists

13

Classification Methods

- Linear Predictor (for 2 classes)
  $L(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_fx_f$
  is a weighted combination of important features to which a classification threshold is applied
  - Examples: Linear discriminant analysis, compound covariate predictor, weighted voting method, support vector machines with inner product kernel, perceptrons, naïve Bayes MVN mixture classifier
- Distance-based
  - Examples: Nearest neighbor, nearest centroid
- Generalizable to > 2 classes

(Simon, *Journal of Clinical Oncology* 2005)

14

Choice of Classification Approach

- Comparative studies of class prediction methods (e.g., Dudoit et al, 2002) have shown simpler methods (LDA, NN) perform as well or better than more complex methods on very high-dimensional marker data (e.g. gene expression microarray)

15

Building a Classifier: Sample Size Considerations for "Training Data"

- Sample size = number of cases, NOT number of features (e.g., genes, proteins) measured
- Sample size determination for training set
  - Large enough to find sufficient number of informative features while controlling false positives (Dobbin and Simon, *Biostatistics* 2005; Dobbin et al, *JNCI* 2003)
  - Large enough so that expected accuracy of resulting classifier is within some tolerance of true accuracy (Dobbin and Simon, *Biostatistics* 2007; Dobbin, Zhao and Simon, *Clin Cancer Res* 2008)
  - Few dozen to few hundred cases required depending on difficulty of prediction problem

16

Quantifying "How good is the classifier?"

- Estimate percent correct classifications ("classification accuracy")
- Survival differences or hazard ratios associated with classification (or with continuous risk score) of sufficient magnitude to be clinically meaningful
- Value added beyond standard clinico-pathologic factors

17

Classification: Avoiding Pitfalls

- When number of potential features is much larger than the number of cases, can always fit a classifier to have 100% prediction accuracy on data set used to build it
- Estimating accuracy by "plugging in" data used to build a classifier results in highly biased estimates of prediction accuracy (re-substitution estimate)
- Internal and external validation of classifier are essential

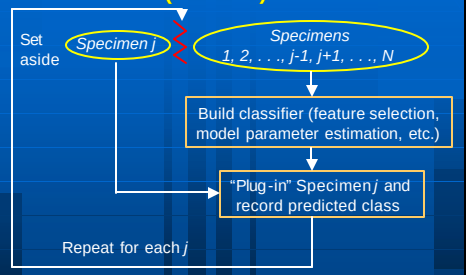
18

Validation Approaches

- Internal: within-sample validation
  - Cross-validation
    (leave-one-out, split-sample, k-fold, etc.)
  - Bootstrap and other resampling methods
  - See Molinaro et al (*Bioinformatics* 2005) for comparison of methods
- External: independent-sample validation

19

Leave-one-out cross-validation (LOOCV)



ALL steps, *including feature selection*, must be included in the cross-validation loop

20

Limitations of Within-Sample Validation

- Frequently performed incorrectly
  - Improper cross-validation (e.g., not including feature selection)
  - Special statistical inference procedures required (Lusa et al, *Statistics in Medicine* 2007; Jiang et al, *Stat Appl Genetics and Mol Biol* 2008)
- Large variance in estimated accuracy and effect sizes
- Doesn't protect against biases due to selective inclusion/exclusion of samples
- Built-in biases? (e.g., lab batch, specimen handling, etc.)

21

Review of Microarray Studies Examining Associations With Cancer Clinical Outcome

(Dupuy and Simon, *JNCI* 2007)

- Detailed account of 42 studies published in 2004 (journals with impact > 6)
- 21/42 studies contained at least one of 3 basic flaws
  - Unstated, unclear, or inadequate multiple testing control
  - Claim of correlation between clusters and clinical outcome after clustering using genes selected for association with outcome
  - Incorrect cross-validation procedure resulting in biased estimation of prediction accuracy

22

There is no substitute for a well-designed, COMPLETELY INDEPENDENT validation study.

23

Steps to Validate Clinical Utility

- Achieve acceptable reproducibility of classification or score
  - Stringent component-wise reproducibility might not be necessary
  - Reference lab versus multiple labs
- COMPLETELY specify
  - Specimen acquisition and handling realistic for clinical use
  - Assay platform (e.g., reagents, chip, equipment)
  - Technical protocol, including quality criteria
  - Data pre-processing
  - Form of classifier or risk score, including cutpoints

24

Steps to Validate Clinical Utility

- Design prospective study
  - Patients representative of target population (e.g., age, stage)
  - Specific treatment context
  - Adequate sample size
- Pre-planned analysis to establish fitness for intended clinical use
  - Clinical outcome measure (e.g., overall survival, distant disease-free survival, tumor response)
  - Performance metrics
    - Percent accuracy
    - Survival curve separation

25

Summary

- Considerable investment of time and resources
- Expertise required: clinical, laboratory, biology, statistics, computational
- Attention to clinical feasibility and affordability
- Clinical impact must be sufficiently high!

26

Acknowledgements

- Richard Simon
- Kevin Dobbin
- Lara Lusa
- Members of the Biometric Research Branch at NCI
- Members of the Cancer Diagnosis Program at NCI

27